Grounding in social media: An approach to building a chit-chat dialogue model

Ritvik Choudhary
Waseda University
ritvik@fuji.waseda.jp

Daisuke Kawahara
Waseda University
dkw@waseda.jp

Abstract

Building open-domain dialogue systems capable of rich human-like conversational ability is one of the fundamental challenges in language generation. However, even with recent advancements in the field, existing open-domain generative models fail to capture and utilize external knowledge, leading to repetitive or generic responses to unseen utterances. Current work on knowledge-grounded dialogue generation primarily focuses on persona incorporation or searching a fact-based structured knowledge source such as Wikipedia. Our method takes a broader and simpler approach, which aims to improve the raw conversation ability of the system by mimicking the human response behavior through casual interactions found on social media. Utilizing a joint retriever-generator setup, the model queries a large set of filtered comment data from Reddit to act as additional context for the seq2seq generator. Automatic and human evaluations on open-domain dialogue datasets demonstrate the effectiveness of our approach.

1 Introduction

Humans have long wanted to talk with the machine and have them comprehend and generate natural language. The task of chit-chat dialogue response generation can be described as one of the major goals in natural language processing. As such, there has been considerable interest in the sub-field of open-domain dialogue models.

Nevertheless, the existing dialogue response generation models still suffer from some very fundamental problems: lack of interesting (“Ok”, “I see", etc.) or uninformative responses (“I don’t know") (Li et al., 2016a, Shao et al., 2017, Ghazvininejad et al., 2017). The primary cause for this is that, unlike humans, the models do not have access to knowledge, experience about out-of-domain topics or human conversational habits and hence can only produce limited unengaging generic responses.

Recent work has proposed considering additional context information such as multi-turn conversational history (Zhang et al., 2018), persona (Li et al., 2016b) or a fact-based knowledge base (Dinan et al., 2019). Among these, our work approaches this problem from a more general standpoint of improving the raw conversational ability of generative models. We attempt this by taking inspiration from how humans learn to converse, i.e., through mimicking social interactions. Applying this in the context of dialogue models, we use a human-readable external knowledge base consisting solely of unstructured **s**ocial **m**edia **i**nteractions (hereinafter referred to as SMIkb), which tends to include a more diverse language structure and hence improve generated responses.

For our approach, we jointly train a generator-retriever model where the retriever searches through pre-indexed SMIkb and feeds the related information together with the input utterance to the generative seq2seq model, allowing for additional context at the time of generation.

In particular, we utilize the Dense Passage Retriever proposed by Karpukhin et al. (2020) on top of BART (Lewis et al., 2020a) as our generational model trained on a mix of open-domain dialogue datasets, together with a collection of Reddit submissions and comments as our main source of social interactions. Experiments showed that our approach outperformed the existing vanilla seq2seq baseline (BART) across all of the automatic and human evaluation metrics. By making use of interactions grounded in social media, the generated responses were not only more engaging but were also shown to be much more relevant and natural, thus establishing the effectiveness of our approach.

2 Related Work

Dialogue Systems In recent years, major breakthroughs beginning with the Transformer (Vaswani et al., 2017) and BERT (Devlin et al., 2019) have

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 9 - 15
July 10-15, 2022 ©2022 Association for Computational Linguistics

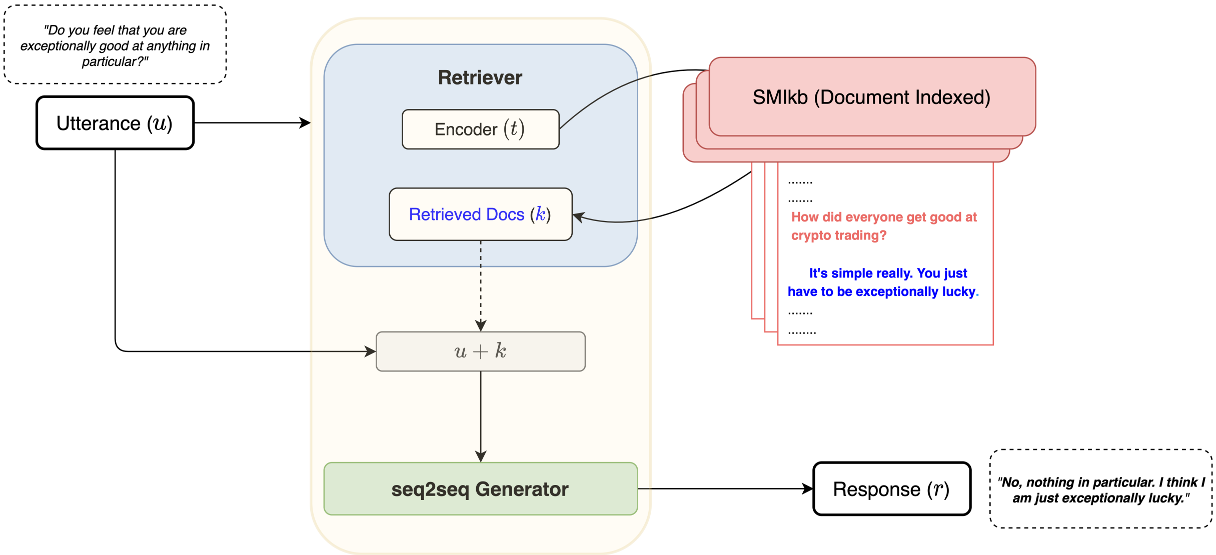

Figure 1: Our proposed dialogue response generation approach grounded in SMIkb through a jointly trained retriever-seq2seq generator setup. Utterance u is encoded and matched against titles (in red) where the respective comments (k, in blue) are retrieved from the SMIkb. These act as an additional context for the generator to generate the final dialogue response r.

quickly shifted the landscape of modern NLP research. These were shortly followed by autoregressive seq2seq models (T5 (Raffel et al., 2020), BART) that significantly improved performance on generation-based tasks such as dialogue systems. We adopt the widely accessible BART as our strong baseline.

Knowledge-based Conversational Models Incorporating additional context or external information into existing models has been a field of much interest lately. Persona-chat (Zhang et al., 2018) or Empathetic Dialogues (Rashkin et al., 2019) take into account persona or empathetic information. Furthermore, advancements making use of knowledge bases in the area of open-domain dialogue systems have become increasingly common (Ghazvininejad et al., 2017; Dinan et al., 2019). The closest work to ours, in terms of including a retrieval step for dialogue generation, is Weston et al. (2018), which proposed an approach involving pre-training the retriever and generating only over the candidates retrieved in advance from the training set. More recently Roller et al. (2021) also tested retrieval-based dialogue generation. However, similar to Weston et al. (2018), they utilized a retrieval model that was kept fixed during training. Our work meanwhile follows a different direction that does not require pre-training of the retriever but fine-tunes it along with the generator to retrieve over a much larger knowledge base of interactions at generation time.

We would also like to mention Shuster et al. (2021), which investigates factual hallucination in dialogue retrieval-generation models with a fact-based knowledge base such as Wikipedia. Our work takes a more generalized approach, focusing solely on improving the raw conversational ability of dialogue models. Instead of factual accuracy, we propose a simple approach for generating an engaging conversation grounded in unstructured social media interactions.

3 Proposed Approach

In this section, we discuss our approach to introducing social media interactions as an external knowledge base (SMIkb) to ground in for more natural and human-like response generation. We begin with formulating the task of dialogue generation and then proceed to explain our joint retriever-generator model as the proposed setup for utilizing the aforementioned unstructured data source. Note that in this work, we primarily focus on response generation for single-turn dialogues or dialogues. We decided that other settings such as a multi-turn case were best addressed in future work.

3.1 Task Formulation

Our task of response generation grounded in external knowledge can be formulated as training a model to predict a response $\mathbf{r} = (r_1, r_2, ..., r_m)$ of

m words when given an input utterance $\mathbf{u}$ and a set of documents $\mathcal{D}$ that might contain relevant knowledge. We define our goal as to allow the model to learn the parameters such that when given an input utterance $\mathbf{u}$ and a knowledge base $\mathcal{D}$, the model can generate a response $\mathbf{r}$ following the probability $p(r_i|\mathbf{u}, \mathbf{r}_{<i}, \mathcal{D}; \theta)$, where θ refers to the parameters of the model.

3.2 Model

Inspired by recent advances in retrieval assisted QA (Guu et al., 2020; Lewis et al., 2020b), we adopt a simple joint retriever-generator setup to the task of dialogue generation. Concretely, we utilize BART, a seq2seq model pre-trained on a denoising objective, as our generative model along with the pre-trained neural Dense Passage Retriever (DPR) (Karpukhin et al., 2020) as the retriever of choice. DPR is a highly efficient neural retriever pre-trained for retrieving the top-$\boldsymbol{k}$ similar documents to an input query $\boldsymbol{u}$. It executes this by encoding both the query and the entire knowledge base through independent BERT-based encoders (as $\boldsymbol{t}$). Furthermore, we follow Karpukhin et al. (2020) to build an offline searchable dense vector index of these embeddings for our SMIkb using the FAISS (Johnson et al., 2017) library for faster lookup. An overview of our architecture is shown in Figure 1. Application of our model to dialogue response generation can be formulated as a two-step process: (1) the retriever searching top-k documents from the pre-indexed interaction knowledge base, relevant to the input utterance, and (2) the generator predicting the response to the previous utterance along with the retrieved context.

Following the notion set in Section 3.1, the probability of generating the response $\mathbf{r}$ given the utterance $\mathbf{u}$ and each of the top-k documents d_j from the knowledge base $\mathcal{D}$ can be defined as

$$p(\mathbf{r}|\mathbf{u};\theta,\lambda) = \textstyle\sum_j^k p_\lambda(d_j|\mathbf{u};\lambda) \prod_i p_\theta(r_i|\mathbf{u}, \mathbf{r}_{<i}, d_j;\theta), \tag{1}$$

where θ and λ are parameters for the generator and retriever, respectively. They are both fine-tuned jointly in an end-to-end fashion, with the retriever providing additional context that is concatenated together with the input at the time of generation. As there is no "correct" document source in the knowledge base, we consider it to be a latent variable. Therefore, during decoding we marginalize these probabilities over all the retrieved documents to return the most probable (best) response using beam search.

Dataset	Total (turns)	Train	Valid	Test
DailyDialog	76,743	53,721	11,511	11,511
DailyDialog++	39,913	27,939	5,987	5,987
Cornell Movie-Dialogs	221,088	154,762	33,163	33,163
Reddit (pseudo extracted)	200,000	140,000	30,000	30,000

Table 1: Overview of datasets in use.

4 Experiments

We evaluate our model together with various external knowledge datasets on a mixture of open-domain dialogue datasets. The results are then compared with two BART-based baselines.

4.1 SMIkb

Aiming to improve the raw communication ability of dialogue systems by mimicking human response behavior, we built our external knowledge base of unstructured social media interactions (SMIkb). It comprises of entries from top thread titles and their top 100 comments from Reddit, an American social news aggregation and discussion site, throughout 2020 (January-November). A total of 1.6 million entries were first scraped through the open-sourced Pushshift API (Baumgartner et al., 2020) of which a random selection of 600,000 (due to memory limitations) makes up our SMIkb. A snapshot of the same is shared in Table 5.

Furthermore, to verify the effectiveness of using a conversational knowledge base like Reddit, we compared ours to a pure Wikipedia knowledge base (ref. "Wiki") of the same size (random sample of 600k entries) containing the wiki page title and the leading 100 words. Additionally, we also tested a 1:1 combination of the above two bases (ref. "Mix").

4.2 Datasets

We fine-tune our models on a variety of open-domain and scraped dialogue datasets.

Open-domain datasets We use a combination of DailyDialog (Li et al., 2017) and DailyDialog++ (Sai et al., 2020) as high-quality daily life-based dialogue sets. We also consider the Cornell Movie-Dialogs Corpus (Danescu-Niculescu-Mizil and Lee, 2011), which is a corpus of scripts of movie dialogues.

Reddit Furthermore we extract another 200,000 comment pairs from Reddit, distinct from the

Model Setup	Training Data	Knowledge Base (Retrieval)				BLEU-4	Dist-1	Dist-2			
Baseline 1	ODD	None				1.31	0.20	0.96			
Baseline 2	ODD + SMIkb	None				1.05	0.12	0.47			
			$k=3$			$k=5$			$k=7$		
			BLEU-4	Dist-1	Dist-2	BLEU-4	Dist-1	Dist-2	BLEU-4	Dist-1	Dist-2
Ours (SMIkb)	ODD	SMIkb	**9.78**	**2.80**	**16.90**	<u>**10.51**</u>	**5.50**	<u>**26.63**</u>	**10.48**	<u>**5.51**</u>	**26.62**
Ours (Wiki)	ODD	Wiki	6.93	2.57	14.91	7.14	4.94	23.38	7.11	5.02	23.79
Ours (Mix)	ODD	SMIkb + Wiki	6.03	2.45	14.08	6.20	4.71	22.25	6.21	4.71	22.23

Table 2: Automatic evaluation of generated responses across various values of k for top-k document retrieval. The baselines do not have a retrieval step and therefore do not have an effect due to changing k. <u>**bold**</u> refers to the best scores across all k among the generated responses. ODD is the collection of **O**pen-**D**omain **D**atasets from Section 4.2.

Model Setup	Human Eval.		
	Relevance	Engagement	Knowledge
Gold (Test-Data)	3.50	3.33	3.47
Baseline 1	2.82	2.35	3.00
Baseline 2	3.03	3.02	2.89
Ours (SMIkb)	**3.84**	3.75	3.60
Ours (Wiki)	3.40	3.75	**3.76**
Ours (Mix)	3.62	**3.80**	3.71

Table 3: Human evaluation of responses for the best $k = 5$.

SMIkb, to act as a pseudo dialogue dataset to supplement our knowledge base.

An overview of the datasets is listed in Table 1.

4.3 Experimental Setup

Implementation Details Our joint retriever-generator model consists of a pre-trained Dense Passage Retriever and BART-large (24 layers, 406M), which are later fine-tuned together on SMIkb and dialogue datasets. The model is trained mostly with the default parameters, batch size of 1, and an initial learning rate of 3×10^{-5}. We further experiment with various values of k for our top-k document retrieval, while beam search with size of 5 is used as our response decoding strategy.

Baseline We consider two strong baselines based on a vanilla BART-large with no retriever to investigate the effectiveness of our approach. The first is fine-tuned solely on the datasets mentioned in Section 4.2 (ref. "Baseline 1") with no SMIkb. Next to confirm the effectiveness of our providing external data through our retriever-generator setup, we merge the entire SMIkb interactions into our training data, and simply fine-tune the vanilla model on this new extended set. (ref. "Baseline 2"). Note that although we choose BART as our generator and baseline for its size and relative ease in training, our proposed SMIkb based modeling setup could possibly also be extended to larger models.

4.4 Evaluation

To measure the impact of social media interactions, the generated responses were evaluated through both automatic and human evaluations. The results are compiled in Tables 2 and 3.

Automatic We perform a series of automatic evaluations on the generated responses. Starting with the BLEU (Papineni et al., 2002) score for relevancy, we also calculate Distinct-N (Li et al., 2016a) for the amount of diversity introduced.

Human It has been widely reported that automatic metrics do not sufficiently evaluate the actual quality of the generated outputs (Liu et al., 2016). Thus, we additionally performed human evaluation of the responses with the highest BLEU ($k = 5$) through Amazon Mechanical Turk, on the following three metrics: *Relevance*, whether the response is relevant to the utterance; *Engagement*, whether the response makes the conversation engaging; and *Knowledge*, whether the response seems knowledgeable or sensible. The evaluators were asked to score 100 responses selected at random from the test set, on a scale of 1-5. Each response was scored by 7 different evaluators, and their average was calculated. We selected our evaluators to be English speakers from the United States with an approval rate of over 90%.

4.5 Discussion

First, with automatic evaluation, we observe that our method of introducing social interactions through a retriever at generation time maintains task performance and allows for a more diverse set of responses, as shown with an increase in all of the

Input Utterance	SMIkb	Wiki	Mix	Baseline
Here comes the birthday cake.	Wow, what a lovely cake. I'll have a piece.	Wow, what a beautiful cake. It must have tasted as good as I expected.	Really? Wow. Happy birthday to you! You look so healthy.	Thank you.
I'd better go. I gotta get up early. Go over to my parents'.	What are you going to do for breakfast?	You mean you don't wanna go with me?	Why don't you stay at my place? I'll cook dinner.	I don't think so.
Something's burning?	Oh my goodness, it's on fire!	Oh my goodness, it ' s burning!	Oh, my bread! It's burning!	No, it's not.
Would you mind if we just called it an early night?	Even I couldn't. I have work tomorrow.	I'd love to. But I've got to work tomorrow.	What are you talking about? We just ate?	I don't think so.
Isn't it wonderful walking here?	What do you mean? Was it bad?	Not at all. The air is fresh, the walk is easy and the sunshine is kind.	Yes, you can see the whole city from here.	Yes, it's very beautiful.

Table 4: A comparison of responses generated from various sources.

SMIkb	
title	text
LPT: If you borrow something like a tool or a generator from someone, return it in BETTER shape than you got it.	My dad always said that returning something in the same condition you received it is the absolute bare minimum.
SoftBank Nears $40 Billion Deal to Sell Arm Holdings to Nvidia	Nvidia is priced decentlyfor what they offer.
Apple to Give Employees Paid Time Off to Vote in U.S. Election	This exactly. A large majority of disenfranchised communities work jobs that don't observe federal holidays.
Apple may be working on a foldable iPhone	I can confirm that Apple would be stupid to not be working on one. Whether they ever release one is up for debate, but they're definitely working on one.

Table 5: Snapshot of SMIkb.

measured metrics over both the baselines. Moreover, our Reddit-based SMIkb model outperformed other combinations, confirming an increase in response quality. Furthermore, experiments comparing the effect of top-k on generation quality showed $k = 5$ as the sweet spot with the highest BLEU, which was then evaluated further.

Similar to the automatic evaluation, all of the proposed combinations showed improvements over the baseline in human evaluation. In line with our initial hypothesis, the SMIkb model recorded the highest score in terms of conversational relevance. Meanwhile, the mix of SMIkb and Wikipedia seems to find the appropriate balance between conversational ability and "knowledgeability" with the metrics even exceeding the gold score. We hypothesize that this is not just due to high generation quality but also the nature of our task, which might make our model responses more relevant to a worker when assessed in a single-turn context. Overall, the results verified our proposed approach of learning directly from social media interactions leading to large improvements over the baseline in all metrics of interest. Furthermore, significance tests of bootstrap re-sampling showed that the above evaluation score differences between baselines and our models were statistically significant ($p < 0.01$). Examples of the various generated responses across all the models are shared in Table 4.

In addition, we looked at cases where our model struggles or fails at generating a natural response, a select few of which are compiled in Table 6. In some of these particular cases, the baseline generations, although generic, turn out to be more relevant. We believe some of these long responses with unrelated information to be an artifact of our generation model being overly dependent on the

Input Utterance	SMIkb	Wiki	Mix	Baseline
Who was the conductor?	Oh speaking of the conductor, he was indeed a genius. A descendent of Chinese.	I don't remember the name.	Oh, speaking of the conductor, he did an excellent job.	I know.
Oh god you scared me!	Ahahah! We need a fitness instructor.	I'm sorry.	Sorry. I was trying to scare you.	Is something wrong?

Table 6: Selected examples where the proposed model struggles or fails at generating a relevant response.

knowledge base. While social media may simulate human-like conversations in a large variety of situations, it is still far from being a perfect stand-in for real-life dialogue. Therefore, our future work in this direction should look at not only the quality and scope of the knowledge base, but also consider selecting *when* to ground and make use of the said knowledge during response generation.

5 Conclusion

We aimed to improve the raw conversational ability of dialogue systems by grounding the responses in much more human-like social media interactions. Our approach involved a neural retriever-seq2seq generator model fine-tuned jointly, where relevant knowledge was retrieved at the time of generation to assist a more engaging and natural dialogue response. Our experiments showed significant improvements with both automatic and human evaluation metrics ranking the SMIkb-grounded replies to be much more diverse, engaging, and relevant.

Acknowledgements

This work was supported by a joint research grant from LINE Corporation.

References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 830–839. AAAI Press.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. *arXiv preprint arXiv:1702.01932*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b.

Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219, Copenhagen, Denmark. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92, Brussels, Belgium. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.